DNA Dinucleotide Evolution in Humans: Fitting Theory to Facts

Alexander Renwick, Leslea Davison, Heidi Spratt, J. Patrick King and Marek Kimmel

Department of Statistics, Rice University, Houston, Texas 77251

Manuscript received December 5, 2000
Accepted for publication July 15, 2001

ABSTRACT

We examine length distributions of ~6000 human dinucleotide microsatellite loci, representing chromosomes 1–22, from the GDB database. Under the stepwise mutation model, results from theory and simulation are compared with the empirical data. In both constant and expanding population scenarios, a simple single-step model with parameters chosen to account for the observed variance of microsatellite lengths produces results inconsistent with the observed heterozygosity and the dispersion of length skewness. Complicating the model by allowing a variable mutation rate accounts for the homozygosity, and introducing a small probability of a large mutation step accounts for the dispersion in skewnesses. We discuss these results in light of the long-term evolution of microsatellites.

MICROSATELLITES are regions of DNA where a short (2–6 bp) motif is repeated a number of times (e.g., . . . ATATATATAT . . . ). They are ubiquitous in that they are found in the genomes of all living organisms. They are also highly polymorphic; the number of repeats varies between individuals. Also, they are easily assayed; the repeat lengths of known loci can be assessed from a small tissue sample.

Polymorphism of microsatellites reflects their high mutation rate, on the order of $5 \times 10^{-4}$ up to $5 \times 10^{-3}$ per generation. Mutation rate is thought to depend, among others, on the length of repeat motif (Weber and Wong 1993; Chakraborty et al. 1997). Mutations are thought to occur by two possible mechanisms: Replication slippage occurs when the DNA polymerase enzyme “slips” during replication. Unequal crossing over occurs during chromosomal recombination when the site of recombination is located within a microsatellite but the DNA strands are misaligned.

The uses of microsatellites include forensics, gene mapping, and evolutionary studies. Because of the large number of alleles, individuals can be uniquely identified by their allele status at several loci (Chakraborty and Jin 1993). The same property helps resolve haplotypes in family linkage studies (Dib et al. 1996). The application of most interest in this article is in evolutionary studies. Phylogenetic relationships and/or demographic events may be inferred from population level repeat length distributions (Mountain and Cavalli-Sforza 1997; DiRienzo et al. 1998; Kimmel et al. 1998; Reich and Goldstein 1998; Reich et al. 1999).

To draw correct inference from microsatellite statistics, it is of importance to understand the evolutionary dynamics of microsatellites. Each particular microsatellite locus finds itself under the action of mutation and drift. In addition, although it is possible to restrict the study to microsatellites in noncoding regions of the genome, linkage disequilibrium may cause them to be associated with other loci, possibly under selection. Another force of importance is demography. Evolution of microsatellites, like that of any other loci, may be influenced by past population expansions, bottlenecks, and migrations (Kimmel et al. 1998).

The usual model invoked in the context of microsatellite loci is the stepwise mutation model (SMM; see Kimmel and Chakraborty 1996 for a review of literature), in which the only form of allelic change by mutation is an extension or contraction in the number of repeats at the locus. The most common version of the SMM is the single-step SMM (or SSMM), in which the extensions and contractions are never longer than one repeat. Long-term consequences of using the SMM instead of a more general model with multiple expansions/contractions possible were explored by several authors and found to be significant.

One way to understand the evolutionary dynamics is to use mathematical modeling. In this approach, predictions of a mathematical model are related to data-based statistics. Such comparisons allow limiting the set of parameters of the model to those that fit the data. In this way, it can be determined whether a given mechanism plays a role in the evolutionary dynamics.

In this article, we focus on length distributions of human dinucleotide repeat microsatellites, obtained from a publicly accessible resource. We construct a family of models of dinucleotide evolution and identify important parameters for models of dinucleotide mutation, including the distinction between single- and multiple-step SMM. We compare model predictions with...
empirical data and discuss implications of our findings for dinucleotide evolution.

DATA

The Genome Database (GDB), established at Johns Hopkins University in 1990, is the official central repository for genomic mapping data resulting from the Human Genome Initiative. In support of this project, GDB stores and curates data generated worldwide by researchers engaged in the mapping effort of the Human Genome Project (HGP). The current version, GDB 6.0, is accessible through the Internet site http://gdbwww.gdb.org/gdb/gdbtop.html. To collect data we used the GDB 5.6 version, which was phased out recently.

For the purpose of this study, we collected data on nongene-associated dinucleotide loci using the GDB 5.6 polymorphism query. The data we extracted include the following: loci names, allele sizes in kilobases, relative allele frequencies, the number of chromosomes sampled, and the literature reference or names of the researchers contributing data. Data collected by us cover chromosomes 1–22, all autosomal chromosomes, jointly covering 3.2 × 10⁹ bp of the DNA (source: Science Maps and Data at the http://www.chlc.org/ScienceData.html/Internet site). This is most likely one of the largest data samples ever compiled in this fashion. At the time our data were collected from the GDB (1998 and 1999), the database included mostly Caucasian (CEPH) polymorphisms of dinucleotides.

The data, culled from the online Genome Database, consist of 5800 dinucleotide repeat loci from 22 chromosomes. In the vast majority of cases, at least 40 chromosomes were sampled per locus. The distribution of the number of loci per chromosome is depicted in Figure 1. We limited ourselves to dinucleotide loci, since quantitative information concerning other microsatellites is scarce in the GDB.

Figure 2 depicts the distributions of the number of alleles and the range of the number of repeats in the dinucleotides sampled. The distribution of the number of alleles is narrower than the distribution of the number of repeats, since, for any given locus, the number of alleles is frequently less than the range of repeat counts might indicate; i.e., within a given range of repeat counts, usually some alleles are missing. However, the shape of the two distributions is remarkably similar. For each locus, heterozygosity (H), variance of repeat count (V), and skewness of the distribution of repeat count (S) were estimated according to the expressions

\[ \hat{H} = \frac{n}{n-1} \left(1 - \sum_{k=1}^{K} p_k^2\right), \]

\[ \hat{V} = \frac{n}{n-1} \left(\sum_{k=1}^{K} p_k (x_k - \bar{x})^2\right), \]

and

\[ \hat{S} = \frac{\sum_{k=1}^{K} p_k (x_k - \bar{x})^3}{\hat{V}}, \]

where \( \bar{x} = \frac{\sum_{k=1}^{K} p_k x_k} {p_k} \), \( x_k \) is the number of repeats in the \( k \)th allele, \( p_k \) is the relative frequency of the \( k \)th allele, \( K \) is the number of alleles observed at the locus, and \( n \) is the number of chromosomes in the sample. Figure 3 depicts scatter plots of heterozygosity and skewness against variance of repeat count. Our definition of skewness differs slightly from the standard one, which involves \( \hat{V}^{3/2} \) in the denominator. Our definition is convenient, since the sample variance of \( \hat{V}(\hat{S}) \) of \( \hat{S} \) from the data is very close to the mean \( \bar{V} \), both calculated over all the loci available. In addition, simulated mean \( \bar{V} \) and \( \hat{V}(\hat{S}) \) behave very similarly to each other. The mean heterozygosity in the sample is equal to \( \hat{H} = 0.68 \), the mean variance of repeat count is equal to \( \hat{V} = 5.4 \), and the sample variance of estimated skewness is equal to \( \hat{V}(\hat{S}) = 5.5 \).

MODEL OF EVOLUTION

We assume the microsatellite loci are neutral with respect to natural selection. Mutation is modeled as occurring in steps: Each mutation changes the length of the existing allele by an integer number of repeat motifs. Drift is modeled by the coalescent approximation of the Fisher-Wright process (Tavare 1995). The models and simulation tools are as described in detail in King et al. (2000).

SMM: In this model it is assumed that mutation occurs with frequency \( \nu \) for all loci and it changes allele length by a random number of repeats, i.e., that the change of allele size \( X \) by mutation has the form

\[ X \rightarrow X + U, \]
where $U$ is an integer-valued random. In the case of the SSM model,

$$\Pr[U = -1] = \Pr[U = 1] = \frac{1}{2}.$$  

We also use the version of the SMM model with an admixture of larger mutation steps (of size $U = u \geq 2$); i.e.,

$$\Pr[U = -1] = \Pr[U = 1] = \frac{1 - p}{2},$$

$$\Pr[U = -u] = \Pr[U = u] = \frac{b}{2}.$$  

**Genetic drift and demographic change:** The evolution of a microsatellite locus is also shaped by drift and demography. The genetic drift (the loss of alleles through random sampling of genotypes of new individuals from the gamete pool) acts with strength inversely proportional to the effective population size. This effect determines the distribution of the branch lengths in the genealogy of the sampled locus. For a sample of $n$ individuals, the genealogy may be partitioned according to the times $T_k$, $k = 2, 3, \ldots, n$, where $T_k$ denotes the time for which the sample represents $k$ distinct lineages. In the case of constant population size, it is known that the coalescence times $T_k$ are independent exponentially distributed random variables, each with parameter $(\frac{1}{2})/(2N)$, where $N$ is the number of diploid individuals in the population.
Under constant population size, the most ancient coalescence times tend to be long relative to branches of the tree associated with more recent bifurcations. For this reason, coalescent trees in constant populations frequently exhibit two or three clusters at the tips of the tree connected with the root by a few long branches. Mutations accumulate on these long branches, accounting for much of the allelic variation observed in the sample. Under mutation-drift equilibrium and constant population size, model predictions are determined by the single composite parameter $\theta = 4N\nu$, where $N$ is understood as the effective population size. Specifically, under the SSMM, expected heterozygosity, variance of repeat count, and skewness are given by closed-form expressions

$$H = 1 - (1 + 2\theta)^{-1/2},$$

$$V = \theta/2.$$  

Other statistics can be estimated by simulation of genealogies, under a variety of evolutionary and demographic scenarios.

When the population size varies over time, the above description of the coalescence process must be modified. The times to coalescence are no longer exponentially distributed. Intuitively, the constant coalescence intensity $1/(2N)$ is replaced by the time-dependent coalescence intensity $1/[2N(t)]$, where $N(t)$ gives the population size $t$ generations in the past. Given that there are $k$ lineages represented in the sample at time $t$, the distribution of the time to coalescence is given by

$$P[T_k \geq \tau] = \exp\left[-\left(\frac{k}{2}\right)\int_0^{t+\tau} \frac{ds}{2N(s)}\right], \quad k = 2, \ldots, n.$$
Figure 4.—Predictions of the single-step stepwise mutation model. Constant-population, mutation-drift equilibrium graphs of (a) heterozygosity vs. variance of repeat count, obtained analytically, and (b) variance of skewness vs. variance of repeat count, obtained using simulations. The ⊕ symbols are placed at mean values in GDB data of the corresponding statistics. c and d depict the sensitivity of simulated heterozygosity and variance of skewness (respectively) to changes in the admixture of multiple-step mutational changes, under mutation rates dispersed with a coefficient of variation equal to 1. Longer “clouds” of simulated points (circles) correspond to reference values with no admixture of multiple-step mutational changes. Shorter “clouds” correspond to simulated values with varying admixture of multiple-step mutational changes.

As a result, the distributions of the coalescence times $T_i$ in a population of variable size will be distorted relative to their counterparts in a constant population.

Two growth patterns are of interest here. The first one, so-called “long-neck,” is rapid expansion from previously established mutation-drift equilibrium. Consider, for example, a population originally of size $N_0$, which undergoes a stepwise expansion $t_i$ generations in the past to its current size $N$, where $N \gg N_0$. Looking backward in time, this demographic expansion corresponds to a sudden increase in the coalescence intensity from $1/(2N)$ to $1/(2N_0)$. The effect of this change on the genealogy of a sample of $n$ chromosomes depends on the time since the expansion. If the expansion event is very ancient, even the coalescence times closest to the root will reflect the current population size $2N$. If, however, the growth is sufficiently recent, the lineages in the genealogy at the time of expansion will be subject to the preexpansion coalescence intensity $1/(2N_0)$, and expected coalescence times for these lineages will be much shorter. With high probability then, the most recent common ancestor of the sample is found close to time $t_i$.

The other growth pattern of interest, the so-called “hourglass,” is a rapid expansion following a bottleneck. The prebottleneck population is supposed to maintain mutation-drift equilibrium. Consider, for example, a population originally of size $N_0$, which goes through a bottleneck of size $N_b$, for duration $t_b$, and then it undergoes a stepwise expansion $t_e$ generations in the past, to its current size $N$, where $N \gg N_0$. The effect of this change on the genealogy of a sample of $n$ chromosomes
depends on the time since the expansion and duration of the bottleneck. For a detailed study, see Kimmel et al. (1998).

RESULTS

We carried out a large number of simulations under different variants of the model. We concentrated on three synthetic characteristics: variance of repeat count, heterozygosity, and variance of skewness. The reason to consider the variance of skewness, instead of skewness itself, becomes clear from examination of Figure 3b. In the data, skewness is distributed symmetrically (to a high accuracy) around 0, while its variance changes with the variance of repeat count.

1. We began by simulating the model predictions under the SSMM, mutation-drift equilibrium, and constant population size. In this case, the only parameter varied was $\theta$. The range of values used was $\theta = 1, 2, \ldots, 20$. For each value of $\theta$, 100 simulation runs were carried out, each of them involving 6000 loci, with the assumed sample size of 40 chromosomes. Figure 4, a and b, depicts model predictions, involving scatter of simulated values due to random fluctuations, for the dependence of heterozygosity on vari-
ance of repeat count (a) and for the dependence of variance of skewness on variance of repeat count (b). The corresponding mean values from the sample are depicted in the same graphs. There exists a systematic departure of the model predictions from the observed values, which extends far beyond the level of random fluctuations.

2. Studies in item 1 were repeated assuming a long-neck scenario, with an expansion $t_e = 4000$ generations in the past, from effective population size corresponding to $\theta_0 = 0.5$, to its current size corresponding to $\theta = 10$, and mutation rate $\nu = 5 \times 10^{-3}$ per generation. These values were selected to obtain a simulated value of $\hat{V}$ close to the sample mean of 5.4. Under the long-neck scenario, the discrepancy between variance repeat count and heterozygosity is amplified. However, the discrepancy between variance of skewness and variance of repeat count stays basically unchanged.

The values of effective population sizes, mutation rates, and times from expansion are approximate values, which seem applicable to the demographic history of Caucasian populations (for more details, see KIMMEL et al. 1998 and KING et al. 2000). Values from a region of the parameter space generate the desired variances. However, the general conclusion concerning the discrepancy remains unchanged.

3. Studies in (1) were repeated assuming an hourglass scenario, with an expansion $t_e = 4000$ generations in the past, from a bottleneck of duration $t_b = 2000$ generations (exploratory simulations were carried out for a range of bottleneck durations). The prebottleneck population size corresponds to $\theta_0 = 5$, the bottleneck size to $\theta_b = 0.5$, and the current size to $\theta = 10$, with mutation rate $\nu = 5 \times 10^{-3}$ per generation. Again, these values were selected to obtain a simulated value of $\hat{V}$ close to the sample mean. Under the hourglass scenario, the discrepancy between variance repeat count and heterozygosity is reduced or even reversed. However, the discrepancy between variance of skewness and variance of repeat count again stays basically unchanged.

4. Further, in the framework of constant population and mutation-drift equilibrium scenarios, an array of simulations was performed including the following modifications in the basic model:

a. Sampling $\theta$ values from a lognormal distribution and varying its expected value and coefficient of variation.

b. Assuming an admixture of larger mutation steps (of size $U = u \geq 2$) and varying the probability $1 - p$ of single step.

c. Assuming an admixture of larger mutation steps (with probability $p$) and varying the size $U = u \geq 2$ of the multiple step.

A summary of results of simulations in (3) is depicted in Table 1. As outlined above, the model was constructed by beginning with the one-parameter SSMM and then proceeding to add parameters until the model expectations matched the observed values of $\hat{V}$, $\hat{S}$, and $\hat{H}$ depicted in the top row of Table 1. In the SSMM we set $\theta$ (the one parameter) to a value that matched the expectation of $\hat{V}$ to $\hat{V}$. In this case, model expectations of $\hat{S}$ and $\hat{H}$ were, respectively, too low and too high (Table 1, row 2). Adding variation to the rate of mutation had little effect on the expectations of $\hat{V}$ and $\hat{S}$ while dramatically decreasing the expectation of $\hat{H}$. Thus modeling the mutation rate as drawn from a probability distribution (lognormal, in particular) and adjusting the mean and coefficient of variation allowed the model to account for $\hat{V}$ and $\hat{H}$ while still expecting too small a value of $\hat{S}$ (Table 1, row 3). Allowing for a rare, large-step mutational change in the model had little effect on the expectation of $\hat{H}$ (provided single-step mutations remained predominant) and increased the expectations of $\hat{V}$ and $\hat{S}$.

For large-step sizes ($U$, equal to seven) the effect on $\hat{S}$ was larger than the effect on $\hat{V}$, and so allowing for a rare large-step mutational change produces model expectations that match $\hat{S}$ as well as $\hat{V}$ and $\hat{H}$ (Table 1, row 4).

The final best fit was achieved using values $\theta = 8$, coefficient of variation (c.v.) of $\nu$ equal to 1, $p = 0.015$, and $u = 7$. Figure 4, c and d, depicts the results of varying the proportion $p$ of large-step mutational changes (with c.v. of $\nu$ equal to 1) on the heterozygosity and variance of skewness, for a range of mean values of $\theta$.

Figure 5 depicts the results of a sensitivity study of varying the parameters around their best-fit values. As seen in Figure 5, variance of repeat count and variance of skewness are insensitive to the changes in the coefficient of variation of $\theta$, while heterozygosity decreases with increasing coefficient of variation. Interestingly, this effect is predictable from Jensen’s inequality (BILLINGSLEY 1986) and is not peculiar to

### Table 1

| Observed and modeled average values of variance, skewness, and heterozygosity of human dinucleotide loci under various scenarios of molecular evolution |
|---|---|---|
| $V^*$ | $S^*$ | $H^*$ |
| Observed | 5.4 | 5.5 | 0.68 |
| Modeled under constant population size | 5.4 | 4.1 | 0.80 |
| SSMM* | 5.4 | 4.1 | 0.68 |
| SSMM and mutation rate variability | 5.4 | 4.1 | 0.68 |
| SSMM, mutation rate variability, and large-step admixture | 5.3 | 5.5 | 0.68 |

* $V$, variance.

$S$, skewness.

$H$, heterozygosity.

* SSMM, single-step stepwise mutation model.
the lognormal distribution (see Appendix). Heterozygosity is quite sensitive to the variation in mutation rates across loci. Using a coefficient of variation of \( \theta \) equal to 1, even without an admixture of multiple-step mutation, allows a close match to the empirical heterozygosity distribution (Figure 6).

Varying the values of \( u \) and \( p \) does not change heterozygosity, while it changes both variance of repeat count and variance of skewness. Furthermore, variance of skewness is more sensitive than variance of repeat count. The discrepancy between the single-step model and the data in variance skewness lies in the tails of the distribution of skewness. In the empirical data, more of the mass is in the tails. Adding a small probability of a large mutation step brings the model into a close agreement with the data. Figure 7 depicts the outcome for the best-fit parameters. Finally, it was verified whether the model with variable mutation rate and multiple-step admixture provides a fit to the data under the long-neck and hourglass scenarios. The answer is in the affirmative (Figures 8 and 9), although the fits seem to be slightly worse compared to those assuming constant population. The best-fit values of \( V \), \( S \), and \( H \) were obtained under the following values: long-neck, \( t_e = 4000 \) generations, \( \theta_0 = 0.52 \), \( \theta = 10.4 \), c.v. of \( v \) equal to 0.9,
DISCUSSION

The Fisher-Wright coalescent provides a flexible framework for investigating multiparameter models. We studied model variants involving multistep mutations, variable mutation rate, and changing population size, in addition to parametric studies of the $\theta$ value. The most important conclusion from our study is that the distributions of GDB dinucleotides can be fitted by the SMM model, if a small admixture of multistep mutations is added to the single-step changes.

In our analyses, limited to human dinucleotide loci, it was not necessary to consider either limits on allele size or directionality of mutational changes. In other microsatellites, the patterns seem to be more complicated. For example, both constraints and directionality had to be included by Deka et al. (1999a), to model distributions of human trinucleotides of several types. Similarly, in the context of Y chromosome tetranucleotides, Cooper et al. (1999) needed the assumption of expansion bias to explain out their data. Major asymmetry exists in disease-causing mutations of trinucleotides at loci such as fragile X, myotonic dystrophy, or Huntington’s disease (Richards and Sutherland 1997). Also,
a highly polymorphic CAG repeat locus, ERDA1, on human chromosome 17q21.3, was recently analyzed by Deka et al. (1999b). It has alleles as large as 50–90 repeats apparently without any disease association but with a high intergenerational instability. Paternal transmissions predominantly result in contractions, whereas maternal transmissions predominantly result in expansions.

Swinton and Amos (2001) argue that a moment measure like our skewness $\hat{S}$ may not be sufficiently sensitive as a tool to detect asymmetry of mutational changes. They introduce index ($\alpha_i$) of asymmetry and apply it to GDB dinucleotides. Our simulations using their index indicate that it is very sensitive to a singular type of asymmetry, namely the presence of a mode of the distribution at, or close to, the extreme-size allele. However, these alleles seem to be most sensitive to typing errors. Therefore, even if in simulations the $\alpha_i$ index seems superior to $\hat{S}$, we found it cautious to adhere to $\hat{S}$.

The effects of changing population size observed in our simulations are consistent with previous observations and theoretical work. Indeed, population expansion results in a transient process characterized by growth of both variance and heterozygosity. However, in the long-neck expansion, heterozygosity grows faster than variance (Kimmel et al. 1998). The net effect is that the so-called imbalance index $\beta$, defined as the ratio of the variance-based estimate of $\theta$ to the heterozygosity-based estimate of $\theta$, remains less than its equilibrium value for up to several thousand generations. The opposite effect is present for the hourglass pattern of population change, at least for a long initial period, exceeding...
our $t_i$ (Kimmel et al. 1998). The estimated value of the imbalance index, $\beta = 1.23$, implied by the mean values of variance and heterozygosity in the top row of Table 1, seems consistent with the hourglass scenario, similarly as it was observed in Kimmel et al. (1998) for tetranucleotide-repeat loci. However, following King et al. (2000), this is not significantly different from 1. This is consistent with equally good fits obtained for each of the three demographic scenarios.

The present analysis demonstrates the importance of accounting for mutation rate heterogeneity when interpreting measures of DNA polymorphism. Heterozygosity is strongly affected by a variable mutation rate. As indicated, to fit the data, it was required to consider an admixture of multistep mutations to the single-step model. This was the only way to fit the heavy tails of the allele length skewness distribution. This suggests that the two proposed mutation mechanisms, replication slippage and recombinatorial misalignment, are both at work in creating variability in microsatellite loci. This conclusion is consistent with studies of long-term evolutionary change in microsatellites in other organisms (Zhu et al. 2000), in which major rearrangements were found in addition to single- and multistep contractions and expansions.

**LITERATURE CITED**


Communicating editor: N. Takahata

**APPENDIX: HETEROZYGOSITY UNDER DISTRIBUTED MUTATION RATE**

Under the SSMR model, heterozygosity is equal to $H(\theta) = 1 - (1 + 2\theta)^{-1/2}$, for any given $\theta$. If $\nu$ and consequently $\theta$, is distributed with density $f(\theta)$ and expected value $E(\theta)$, then the expected heterozygosity is equal to $E\theta[H(\theta)] = \int [1 - (1 + 2\theta)^{-1/2}]f(\theta) d\theta$. Since the function $1 - (1 + 2\theta)^{-1/2}$ is concave in $\theta$, we have from the Jensen’s inequality (Billingsley 1986) that $E\theta[H(\theta)] \leq H(E(\theta))]$, which explains the reduction of heterozygosity under distributed mutation rate.